

# Open Source Data Mining mit WEKA/Pentaho



Dr. Alexander K. Seewald



# Was ist WEKA? (1)

## Waikato Environment for Knowledge Analysis

- Benannt nach einem neugierigen flügellosen Vogel, der in Neuseeland heimisch ist und unter Naturschutz steht
- 1000+ Contributors seit 1999, GPL
- Stabilität, Verfügbarkeit und Qualität der Lernalgorithmen weit jenseits von kommerziell verfügbaren Tools
- Sponsor: Pentaho Corp. – in den Top Ten wichtigsten OS Projekten 2007, Infoworld



**Die** weitverbreiteste Data Mining Suite, für Anwendung, Lehre und Forschung

<http://www.cs.waikato.ac.nz/~ml/weka>

# Was ist WEKA? (2)

The screenshot displays three overlapping windows from the WEKA software suite:

- Weka GUI Chooser:** Shows the WEKA logo (a kiwi bird) and the text "WEKA The University of Waikato". It lists applications: Explorer, Experimenter, KnowledgeFlow, and Simple CLI. Below, it states "Waikato Environment for Knowledge Analysis Version 3.6.2 (c) 1999 - 2010 The University of Waikato Hamilton, New Zealand".
- Weka Explorer:** Shows the "Selected attribute" panel for "sepal.length". It displays statistics: Minimum (4.3), Maximum (7.9), Mean (5.843), and StdDev (0.828). Below is a histogram for "Class: class (Nom)" with bars for values 4.3, 6.1, and 7.9, with counts 16, 30, 34, 28, 25, 10, and 7.
- Weka Classifier:** Shows a "Cost/Benefit Analysis" for a Naive Bayes classifier. It includes two plots: "Plot: ThresholdCurve" and "Plot: Cost/Benefit Curve". Below the plots are controls for "Threshold" and "Confusion Matrix". The Confusion Matrix shows 100% classification accuracy. The Cost Matrix shows a total population of 150 and a cost of 66.67.

# Was ist WEKA? (3)

Weka KnowledgeFlow Environment

DataSources | DataSinks | Filters | **Classifiers** | Clusters | Associations | Evaluation | Visualization

CostSensitive Classifier | CVParameter Selection | Dagging | Decorate | END | Ensemble Selection | Filtered Classifier | Grading | Grid Search | Logit Boost | Meta Cost | Multi BoostAB | MultiClassi

Knowledge Flow Layout

Status | Log

Component	Parameters	Time	Status
[KnowledgeFlow]		0:16:0	OK.
ArffLoader		0:0:42	Loading cpu.arff
AttributeSelection	-E "weka.attributeSelection.CfsSu...	0:0:42	Finished.
Discretize	-R first-last	-	INTERRUPTED

# Übersicht

- **Best Moves – Dancing Guide\***
- **Ein Frühwarnsystem für Bot-Netze°**
- **Low-cost Eyetracking\***
- **Watching C. elegans Think°**
- **deFlicker\***

*\* = Video, ° = kompletter Code unter GPL verfügbar*

# Best Moves – Dancing Guide (1)



- Video

# Best Moves – Dancing Guide (2)



- „Berechnet“ den dazupassenden Tanz zur gegebenen (Ball-) Musik
- Seit November für Android, seit Mitte April auch für iPhone
- Analysiert Takt- und Audiosignal-Features (Onset detector histogram, MFCC) direkt am Handy – keine Netzwerkverbindung notwendig
- Online Feedback möglich
- Training mit WEKA basierend auf ca. 600 Testliedern

# Ein Frühwarnsystem für Bot-Netze (1)

## Forschungsprojekt im Bereich IT Security

- Komplementär zum klassischen Spamfiltern
- Vorbeugende Identifizierung und Früherkennung der Ursache von Spam – Bots bzw. Bot-Netze

## Vorgehensweise

- Referenzdaten zu bekannten Bots- und Bot-Netzen
- Trainieren von Lernmodellen zur Erkennung von TCP/IP-Traffic eines bestimmten Bots
- Validierung und Test

*Basiert vollständig auf Open-Source Software; WEKA wird für alle Lernmodelle & spezifische Vorverarbeitung verwendet.*

***Top downloaded journal paper in Q4/2009***



# Ein Frühwarnsystem für Bot-Netze (2)



Verschiedene Farben zeigen Zugriffe durch verschiedene Spambots an. GPL code: <http://botnetz-tracker.seewald.at/>

Hintergrund: [Visible Earth \(NASA\)](#), IP-Positionsbestimmung durch [IP Address Location](#). Spambot Trainingsdaten zur Verfügung gestellt von [Marshal Trace](#).

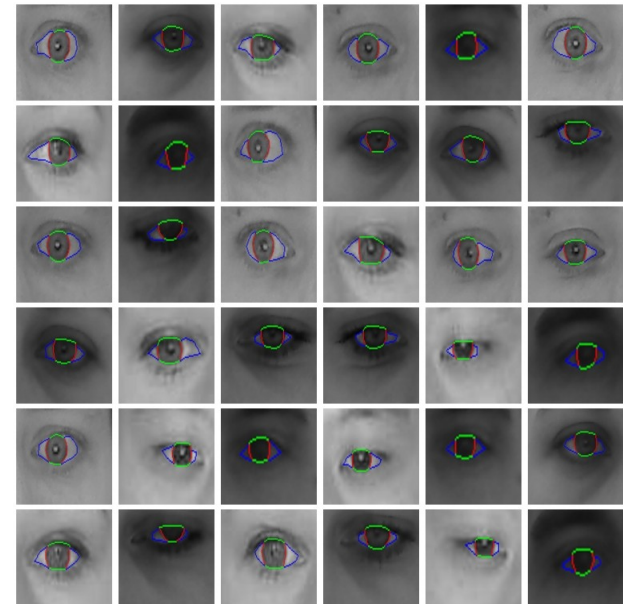
# Low-cost Eyetracking

## Warum?

- teure Eye-Tracking Systeme mit Spezialhardware
- keine Anwendung in gewohnter Umgebung möglich
- keine Analyse von reinem Videomaterial möglich

## Entwicklung eines Low-cost Eyetrackers

- Off-the-shelf USB-Kameras, integrierte Notebook-Kameras, Handycams, ...
- Analyse Videomaterial möglich
- Segmentierung der Augen durch in WEKA trainiertes Modell (drei Klassen: Haut, Sclera, Iris/Pupille)



# Watching C. Elegans Think (1)

## Basic research project in Systems Neuroscience

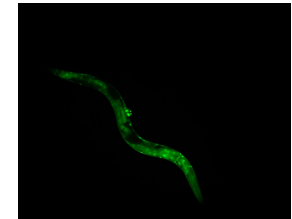
### Four Objectives

- Engineering *Real-time tracking nerve cells*
- Methodological *Validate nervous cell models*
- Holistic *Understand complete N.S.*
- Insight *Better learning algorithms*

Model organism: C. elegans

~ 1000 cells, ~ 300 nerve cells

*Might* be feasible to simulate



# Watching C. Elegans Think (2)

## Results of an automated analysis of C.elegans images (data by Prof. T. Johnson's group)

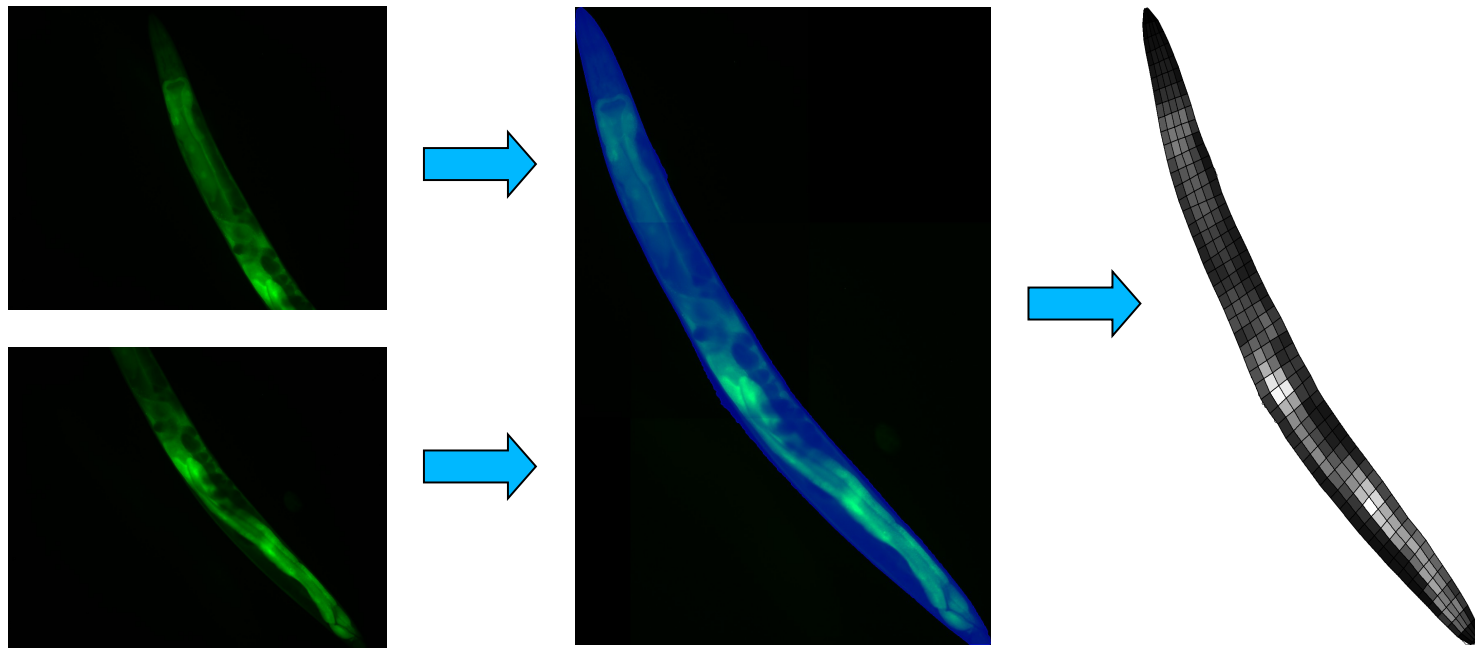


Image processing done via ImageIJ & WEKA

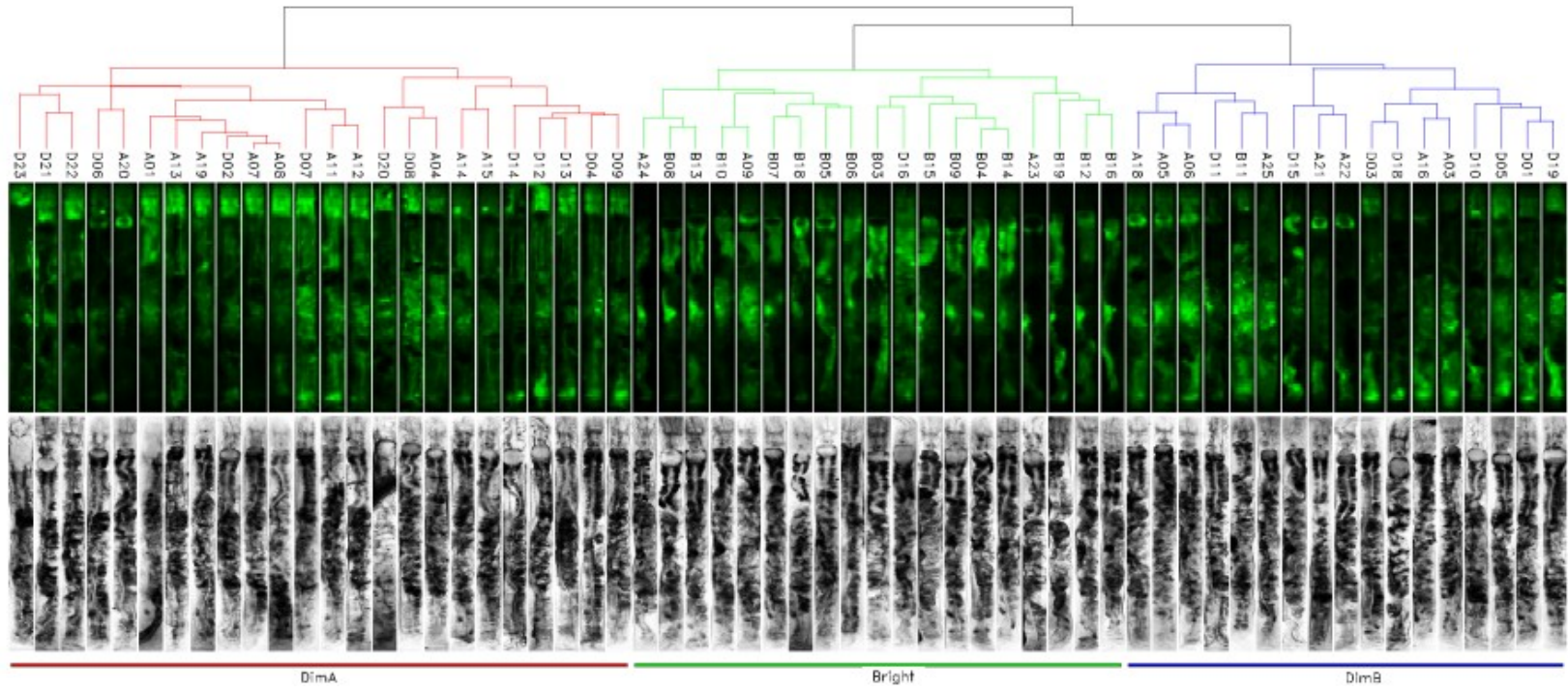
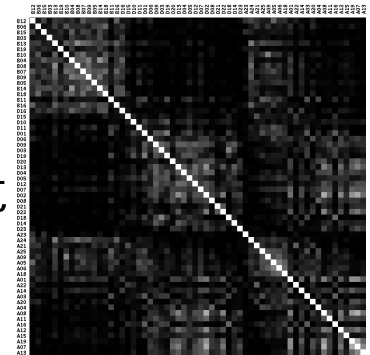
Reduces workload by 80%, paper upcoming

Details & GPL v3 code: <http://elegans.seewald.at/>

# Watching C. Elegans Think (3)

## Some interesting results:

Bright worms live longer than dim worms.  
Even when discounting brightness, bright worms show distinct expression patterns.



# deFlicker



## Warum?

- Weit verbreitete Scheinwerfer flackern auf hoher Frequenz.
- Nachtaufnahmen mit (>150fps) Hochgeschwindigkeitskameras flackern deshalb stark.

## Entwicklung eines Prototypen

- Entfernung des Flackerns in Echtzeit (720p / 50-60Hz)
- Eingesetzt bei UEFA Testspielen, Olympia 2010 (Vancouver), derzeit ausgestellt bei NAB 2010 (Las Vegas)

**Vielen Dank für die Aufmerksamkeit!**

**Für Fragen stehe ich jederzeit  
gerne zu Ihrer Verfügung.**